Integr Mater Manuf Innov (2017) 6:160–171
DOI 10.1007/s40192-017-0094-3

TECHNICAL ARTICLE

Context Aware Machine Learning Approaches for Modeling Elastic Localization in Three-Dimensional Composite Microstructures

Ruoqian Liu[1] · Yuksel C. Yabansu[2] · Zijiang Yang[1] · Alok N. Choudhary[1] · Surya R. Kalidindi[2,3] · Ankit Agrawal[1]

Received: 29 January 2017 / Accepted: 13 March 2017 / Published online: 31 March 2017

Abstract The response of a composite material is the result of a complex interplay between the prevailing mechanics and the heterogenous structure at disparate spatial and temporal scales. Understanding and capturing the multiscale phenomena is critical for materials modeling and can be pursued both by physical simulation-based modeling as well as data-driven machine learning-based modeling. In this work, we build machine learning-based data models as surrogate models for approximating the microscale elastic response as a function of the material microstructure (also called the elastic localization linkage). In building these surrogate models, we particularly focus on understanding the role of contexts, as a link to the higher scale information that most evidently influences and determines the microscale response. As a result of context modeling, we find that machine learning systems with context awareness not only outperform previous best results, but also extend the parallelism of model training so as to maximize the computational efficiency.

Keywords Materials informatics · Machine learning · Elastic localization prediction · Ensemble learning · Context aware modeling

✉ Ankit Agrawal
ankitag@eecs.northwestern.edu

Ruoqian Liu
rll943@eecs.northwestern.edu

Yuksel C. Yabansu
yabansu@gatech.edu

Zijiang Yang
zijiangyang2016@u.northwestern.edu

Alok N. Choudhary
choudhar@eecs.northwestern.edu

Surya R. Kalidindi
surya.kalidindi@me.gatech.edu

[1] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA

[2] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[3] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Introduction

There has been a growing popularity in the use of data mining and machine learning methods in studies of various phenomena in materials science. In particular, parametric models are learned from massive amounts of collected data, either from laboratory experiments or from computational simulations, in order to represent, describe, and approximate process-structure-property (PSP) relationships for materials systems [1–3]. Models built in such manner are often used as surrogate models for the more expensive and/or computationally intensive physically based models (e.g., thermochemical and microstructural evolution models). In contrast to the modeling style in physical models, where we explicitly specify equations, physical constraints, variable spaces to the extreme, data models often free the designers from such specifics. The logical and mathematical formula they use tend to form automatically or semi-automatically, with only the supply of data examples, and specifications of model structures, loss functions, and optimizers, to extract unknown correlations between inputs and outputs. In structure-property modeling of hierarchical materials, the microstructure serves as the input, and the output is usually

a microscale distribution of a field of interest (e.g., stress, strain). Data models offer fast and efficient solutions for the inverse problems posed in materials design. The potential of materials data science approaches has been demonstrated in numerous earlier works [1–28].

The field of machine learning has advanced significantly in recent years, in part due to the prominent progress made in supervised learning [29]. These advances have opened new avenues for stronger collaborations between materials scientists and computer scientists. In these collaborations, it is important to judiciously fuse data-driven modeling with the known or established principles in materials science. It should be recognized that data models are generally designed with computational efficiency as the primary goal. And what they eventually end up learning are almost purely data correlation. They are known to operate in an agnostic manner and function as a black box. On the other hand, physically based models aim to explicitly address the prevailing mechanisms and the complex internal structure of the material system studied. Therefore, it stands to reason that a thoughtful fusion of these approaches can, in principle, allow us to exploit the respective benefits of both approaches. We exploit a fusion design by making sure that it, on the one hand, rightfully depicts the characteristic of materials systems, and on the other hand, stresses the computational parallelism in function optimization and expands the converage of data diversity.

An important attribute of most materials systems employed in advanced technologies is that they exhibit heterogeneous structures at a multitude of hierarchical length scales. Consequently, most materials phenomena of interest are inherently multiscale, and the communication of the salient information between the hierarchical structure scales is the central challenge in any modeling and simulation effort. Historically, the multiscale materials modeling efforts have addressed either homogenization (communication of information from the lower length scale to the higher length scale) [30–33] or localization (communication of information from the higher length scale to the lower length scale) [8, 13, 14, 34–36]. Although both homogenization and localization have been studied extensively in literature using physically based approaches [31–33, 37], recent work has identified the tremendous benefits of fusing these approaches with data-driven approaches [8, 13, 14, 30, 34–36]. However, most of the prior effort has only addressed a limited number of the multiscale features. For example, many of the previous efforts are not readily extendable to high contrast composites (i.e., large differences in the properties of the microscale constituents in the composite).

Other predictive modeling work in materials science domain, whether it is to predict the melting temperatures of binary inorganic compounds [38], the formation energy of ternary compounds [6], the mechanical properties of metal alloys [39], or which crystal structure is likely to form at a certain composition [40, 41], also sees the limitation of learning with a single agent.

In this paper, we design a new data modeling approach that is explicitly hierarchical to be able to take advantage of the multiscale characteristics of a heterogeneous material structure. The design is to be manifested through the idea of context detection, which is a concept used in reinforcement learning and robotics to deal with non-stationary environments [42]. Context detection is defined here as finding the right high-level, low-dimensional, knowledge representation in order to create coherent learning environments. Once different contexts are identified from data, one can build separate models out of each context group. This approach has many similarities with the divide and conquer scheme, which breaks a large, difficult problem into a set of small, simpler problems.

This work examines the advantage of building context aware learning systems by solving an elastic localization problem in high contrast composites. More specifically, we aim to provide a computationally efficient surrogate model, with parameters learned from data, to predict the microscale elastic strain field in any given voxelized three-dimensional (3-D) volume element of a high contrast composite subjected to a prescribed macroscale elastic strain (applied as a periodic boundary condition). We address the multiscale challenge by identifying and representing the higher level data distribution through context detection. In our designed two-stage system, the first stage attempts to find the contexts in data, while the second stage builds context specific learning models. We compare the results from this new data model with benchmarks from previous work [8] and demonstrate that the two layer data modeling scheme provides a viable approach for capturing the elastic localization linkages in high contrast composites.

Methods

Localization: Problem and Data Description

As mentioned in the previous section, multiscale modeling and materials design involves a bi-directional exchange of information between the material length scales (i.e., homogenization and localization). Most efforts have been focused on the homogenization; however, localization is as crucial as homogenization for materials design. Localization denotes the distribution of response field at the lower length scale for a load imposed on the higher length scale. Localization in multiscale materials modeling is depicted in Fig. 1. The target is to find the microscale response field of the microstructure shown. The microstructure is embedded at a macroscale material point. A physically based

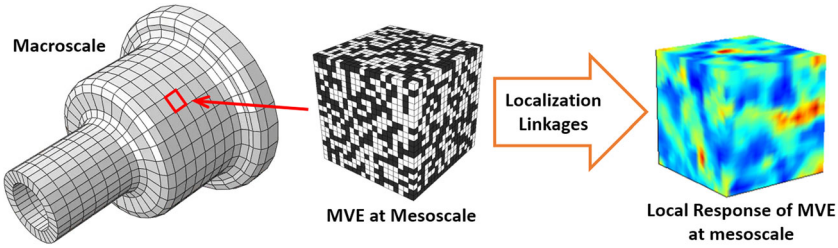

Fig. 1 The depiction of the workflow of localization employed in multiscale materials modeling. The local response of an MVE shown in the middle can be obtained by solving the governing field equations of a physics-based model based on the loading conditions on the location of MVE in macroscale (Color figure online)

model needs to be executed on the microstructure for the boundary conditions extracted from macroscale considerations. Since these models involve solving governing field equations that account for highly complex heterogeneity (at the microscale), they are often computationally cumbersome. A data-driven model offers an opportunity to bypass the physically based approach in a computationally efficient manner, while providing good predictions. Successful extraction and utilization of such data-driven models can dramatically accelerate the multiscale materials modeling and design tasks.

Data-driven models for the localization problem in different material systems subjected to a variety of materials phenomena have been the focus of our prior efforts [13, 14, 34–36, 43, 44]. Most of these efforts were built on model forms suggested by Kroner's expansions obtained from classical statistical continuum theories [45, 46]. In a recent study [8], we explored alternative approaches that were based completely on machine learning techniques that demonstrated significant promise for high contrast composites. This is particularly signficant because Kroner's expansions are known to be applicable to only low to moderate contrast composites [45, 46]. The machine learning-based localization approaches can therefore fill a critical gap, where there do not currently exist versatile and robust solution methodologies that are also computationally low cost.

Elastic localization in a high contrast composite material system serves as a good model system for exploring the emergent machine learning-based data-driven models for localization. In these problems, the goal is to predict the local elastic strain field at the microscale in a two-phase composite material system with constituents that exhibit linear isotropic elastic constitutive response. Because of our interest in high contrast composites, we will assume that the constituents exhibit a contrast of 10 in Young's modulus of microscale constituents with the same Poisson ratio of 0.3.

Two thousand five hundred microscale volume elements (MVE) were generated with 100 distinct volume fraction values. In other words, the entirety of the 2500 MVEs is composed of 100 distinct groups where all 25 MVEs in each group had the same volume fraction. The two distinct phases are distributed in the MVEs randomly. From the set of 25 MVEs generated for each volume fraction group, 15 MVEs were used for the training/calibration, and the other 10 MVEs were used for validation of the linkages. Therefore, a total of 1500 MVEs are used in the calibration of linkages while the remaining 1000 microstructures are used for the assessment of the performance of the linkages. Each MVE has $21 \times 21 \times 21 = 9261$ voxels. In all the MVEs used in this study, each voxel is completely occupied by one of the microscale constituents (i.e., no voxel is occupied by mixtures of the two constituent phases; also called eigen-microstructures [47]). The microscale elastic strain distributions in each MVE were computed using finite element (FE) models executed using the commercial software package, ABAQUS [48]. Linear elastic deformation involves six unique strain values that are represented by ϵ_{ij} where $i, j = 1, 2, 3$ and $\epsilon_{ij} = \epsilon_{ji}$. We will restrict our attention in the present study to the predictions of the strain component, ϵ_{11} (we anticipate that the protocol can be easily repeated for the other strain components). Periodic boundary conditions were imposed on each MVE in such a way that only the average macroscale value of ϵ_{11} was nonzero (i.e., $\langle\epsilon_{11}\rangle \neq 0$ where $\langle\rangle$ represents the macroscale average). Indeed, the boundary conditions described above can be repeated for each of this macroscale six strain components. If this is done, the response field of a volume can be predicted for any arbitrary loading condition using the superposition principle ([49, 50]).

Machine Learning Problem Definition

A discretized microstructure function, m_s^h, is used in this paper to mathematically describe the microstructure in any selected MVE. More precisely, m_s^h denotes the volume fraction of local state h (i.e., black or white phases in Fig. 1) in the spatial voxel of interest, s, where $s = 1, 2, ..., S$. For the case studies shown here, MVEs are generated as eigen-microstructures [47], where each voxel is occupied by a single discrete local state and they have a size of $21 \times 21 \times 21$ resulting in total of $S = 9261$ voxels. In other words, m_s^h

takes values of 0 or 1 depending on the local state occupying each voxel in each MVE. Similar descriptions have been successfully implemented in prior work for microstructure classification [9, 10, 25], microstructure reconstructions [47], and establishing process-structure-property linkages [28, 30, 34, 44].

We reiterate here that the learning problem of interest is to predict the local response p_s (i.e., ϵ_{11} in spatial voxel s) based on the local microstructure m_s^h and its neighborhood information. The combination of information at spatial voxel s and its neighborhood can be organized using a triplet of indices (s, l, and t), described in detail below:

- Neighborhood level l is used to group neighbors based on their distance from the main voxel of interest. More specifically, it includes all voxels in the neighborhood that are at a center-to-center Euclidean distance of $\sqrt{l}$ from the voxel of interest s. For example, the neighborhood level $l = 0$ contains only the voxel s as the distance is 0. For $l = 1$, there are six voxels which are one cell away in both negative and positive reference directions of the three-dimensional coordinate system. As another example, there are no neighbors at $l = 7$ since there are no voxels that corresponds to center-to-center euclidean distance $\sqrt{7}$ away from spatial voxel s.
- Index t enumerates the voxels in a neighborhood level. All voxels in the same neighborhood level have the same importance; thus, they are simply indexed with integers. For instance, voxels at neighbor level $l = 1$ for a selected voxel s can be indexed as $(s, 1, 0)$, $(s, 1, 1)$, $(s, 1, 2)$, $(s, 1, 3)$, $(s, 1, 4)$, and $(s, 1, 5)$. MVEs used in this study are assumed to be periodic and periodic boundary conditions are invoked in all FEM simulations. If the voxel in interest, s is close to the border of MVE, the neigborhood information is selected according to the periodicity in the microstructure.

Following this nomenclature, $m_{s,0,0}$ actually corresponds to microstructure function m_s^h which is the local microstructure information without the neighborhood information. In other words, $m_{s,0,0}$ is considered the focal voxel and p_s is its corresponding local response. In our prior work [8], the main protocol we followed for predictive modeling was composed of two key processes: (i) feature extraction, and (ii) construction of a regression model where the local neighborhood information $m_{s,l,t}$ and the local response p_s were used as input and output, respectively. The utility of each process was evaluated and demonstrated through an ablation test with various data experiments. However, in that work, all data experiments performed predictive modeling in a "global" fashion, where MVEs that span a long spectrum of structural (morphological) characteristics were used in the same model without discrimination. The result of this approach was a single-agent model learned from all available MVEs. The target of this work is to improve the approach presented in [8] by increasing the efficiency in building the model and increasing the accuracy of predicting the linkages.

The concept of this work is to enforce context awareness to the modeling. We define contexts in this scenario as the a priori processing and/or structural information that is responsible for the end property. The means to achieve context awareness is through multi-agent learning. The research on multi-agent learning has been pursued by the machine learning community but was mainly limited to reward-based reinforcement learning problems [51]. Supervised learning problems, as those we encounter in this work, are not systematically addressed. Another related area of research is ensemble learning, where a collective of different classifiers are trained on the same data and the final outcome is formed by an aggregation (usually, an average) of their individual results. However, in an ensemble, the fact that each agent learns from the same data diminishes the diversification impact that supposedly could help with the prediction performance. Although in some cases as random forest classifiers, multiple agents are built from randomized subsets of data, the lack of prior knowledge of each subdivision makes it less applicable to our problem. In our design, we make use of existing information from microstructure and convert it into useful representation of contexts, and consequently have multiple agents that separately learns from each context.

Design of Experiments

The variation of volume fractions in MVE samples first inspired us to construct a divide-and-conquer scheme with multiple learners, each learned from and designated to predict for a specific data class. The key is to divide data samples into multiple classes. The division strategy can be either a hard threshold (e.g., based on values of volume fraction) or a fuzzy statistical boundary. While volume fraction-based division is relatively straighforward, it may not result in the best model. Indeed, in high contrast composites, the mechanical response of different MVEs can be dramatically different even with the same volume fraction. Therefore, it is necessary to seek more generalization by building a learning system that automatically extracts microstructure characteristics, in order to group statistically similar MVEs into the same context.

The basic idea explored in this work is to create a homogeneous modeling environment for each machine learning predictor by assembling MVEs with structural similarity. This two-level learning intends to break down a large-scale prediction problem into multiple subproblems, each having the same or similar setup, but targeting a different facet

of input distribution. In the end, the multiple sub-models learned are selectively and collaboratively applied on new samples, and one or more sub-models can be invoked to contribute to the final prediction.

The outline of such a learning structure is shown in Fig. 2b. Compared with Fig. 2a, which is the framework used in [8], the major advancement is the designed two layers of feature extraction: the macro-layer, and the micro-layer. The macro-layer features are generated to probe MVE-level similarity in structures. Once a representation of similarity is defined, the original set is divided into multiple subsets with low inter-similarity and high intra-similarity. MVEs within the same context group possess higher resemblance with each other, in terms of solely their structure representation, than those from different groups. Micro-layer features, on the other hand, specialize in learning the voxel level characteristics. The influence of various selections of micro-level features has been discussed in [8]. In this paper, we adopt the set of 57 features that gave the best performance in [8]. Features are listed in Table 1. pr_l^h stands for the fraction of voxels with microstructure phase h at neighbor level l. Pr_l^h stands for the fraction of voxels with microstructure phase h *up to* neighbor level l. I_{norm}^h is the normalized impact of all 12 levels of neighbors of phase h. Definitions are provided in Table 2 (further definitions for S_3 and S_9, the symmetry indices, can be found in [8]).

In the approach explored here, the macro features manage the task of "divide" and micro features are in charge of the "conquer" phase of the design. Within each context group (shown in Fig. 2), the same single-agent learning procedure as developed in [8] is followed. In an effort to developing macro features for data-driven strategies, we propose the following data experiments in this paper (M stands for multi-contextual):

- *Ex M1* Use the volume fraction to divide data classes.

Table 1 The input feature set used in our experiments contains 57 voxel-level features. In [8] it is shown to be the best performing set

Rank	Feature
1	$m_{s,0,0}$
2 – 7	$m_{s,1,2}$, $m_{s,1,3}$, $m_{s,1,1}$, $m_{s,1,0}$, $m_{s,1,4}$, $m_{s,1,5}$
8 – 13	$m_{s,2,2}$, $m_{s,2,3}$, $m_{s,2,0}$, $m_{s,2,1}$, $m_{s,4,4}$, $m_{s,2,4}$
14 – 16	pr_1^1, $m_{s,2,6}$, pr_1^0
17	I_{norm}^0
18	Pr_1^1
19 – 20	S_9, S_3
21 – 23	$m_{s,2,8}$, $m_{s,2,5}$, $m_{s,3,3}$
24	Pr_1^0
25 – 30	$m_{s,3,6}$, $m_{s,5,6}$, $m_{s,5,10}$, $m_{s,2,9}$, $m_{s,8,28}$, $m_{s,5,11}$
31 – 36	$m_{s,5,3}$, $m_{s,5,2}$, $m_{s,5,7}$, $m_{s,5,20}$, $m_{s,5,15}$, $m_{s,3,1}$
37 – 42	$m_{s,5,14}$, $m_{s,5,17}$, $m_{s,2,10}$, $m_{s,5,16}$, $m_{s,6,7}$, $m_{s,6,0}$
43	I_{norm}^1
44 – 47	$m_{s,9,5}$, $m_{s,6,2}$, $m_{s,6,12}$, $m_{s,6,5}$
48 – 53	$m_{s,6,3}$, $m_{s,5,21}$, $m_{s,6,1}$, $m_{s,2,11}$, $m_{s,5,19}$, $m_{s,6,6}$
54 – 57	$m_{s,5,18}$, $m_{s,5,23}$, $m_{s,6,9}$, $m_{s,6,4}$

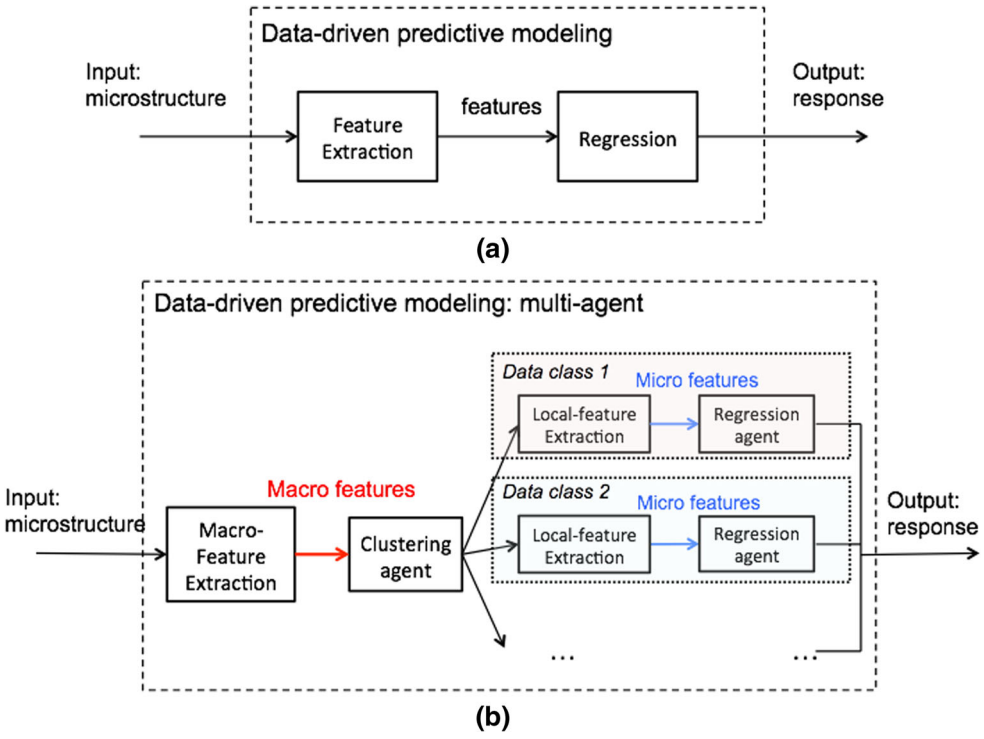

Fig. 2 Flowchart of data-driven predictive modeling. **a** Single agent [8]. **b** Multi-agent (Color figure online)

Table 2 Definition of micro-layer features used in all experiments, with regards to the representation of a focal voxel at s

Symbol	Meaning	Count	Scope
$m_{s,l,t}$	Microstructure value of voxels at a neighbor level l, with index t, of a focal voxel at s $l = 1, \ldots, 12$	179	$binary$, {0, 1}
pr_l^h	Fraction of voxels with microstructure phase h at neighbor level l	24	real, [0,1]
Pr_l^h	Fraction of voxels with microstructure phase h up to neighbor level l	24	real, [0,1]
I^h_{norm}	The normalized impact of all 12 levels of neighbors of phase h $I^h_{\text{norm}} = \sum_{l=1}^{12} T_l \cdot pr_l^h/\sqrt{l} + T_0 \cdot pr_0^h/0.5$	2	real
S_3	3-plane symmetry index	1	real
S_9	9-plane symmetry index	1	real

- *Ex M2* Use designed macroscale microstructure descriptors to divide data classes.
- *Ex M3* Use the characteristic shape of pair correlation functions (PCF) to divide data classes.

Results and Discussion

Dataset Details

A total of 2500 MVEs with varying volume fractions were included in this study. They are evenly distributed in 100 variations of volume fraction (VF) values, from 1.0 to 99.4%. Therefore, 25 MVEs are present in each variation, within which 15 are used for calibration (for feature extraction, model training), and the remaining 10 are used for validation.

The data experiments were carried out on a Linux Red Hat 4.4.7 system with 32 GB memory and Intel Xeon CPU 2.20 GHz. Python-based machine learning library, scikit-learn [52], is used in most implementations (except the M5 model tree which is implemented in a C library). The performance of the models was evaluated by the mean absolute strain error (MASE) e in a MVE, defined as

$$e = \frac{1}{S}\sum_{s=1}^{S}\left|\frac{p_s - \hat{p}_s}{p_{\text{imposed}}}\right| \times 100\% \tag{1}$$

where p_{imposed} denotes the average strain imposed on the MVE, and p_s and $\hat{p}_s$ denote the values of the strain in the voxel s from the FE model and the surrogate model developed in this work, respectively. This metric quantifies the average error for a single MVE microstructure. In the data experiments presented here, we show both individual e for each MVE as well as averaged MASE, $\bar{e}$, over the entire set of 1000 validation MVEs.

In constructing training and testing data for predictive modeling, each voxel in the MVE is examined, represented, and transformed into a data instance consisting of "inputs" and "outputs." Each MVE generates 9261 data samples (this is the number of voxels in each MVE). The complete calibration set hence contains 13,891,500 samples and validation contains 9,261,000 samples.

We term the voxel under examination as the "focal voxel," whose response (average elastic strain in the voxel) is to be predicted. Each voxel in the MVE gets to be the focal voxel once, and when it does, other voxels in its local environment are taken to construct input features for it. By doing this, we are assuming that the response of a focal voxel is strongly influenced by some short-range interactions with neighboring voxels in its local environment. This concept is highly consistent with the concepts of Green's functions utilized extensively in composite theories [45, 46, 53–56].

Context Detection with Volume Fractions

The detection of contexts is carried out with three data experiments. As a baseline, *Ex M1* is designed based on the simplest notion that MVEs are naturally categorized by their volume fractions. 100 variations of volume fractions in the data result in 100 individual contexts and hence 100 branches of prediction systems are built. The testing procedure is straightforward. An incoming MVE will have its volume fraction category easily obtained, and therefore will be handled by the system that is trained with that particular category (or closest to that category). This experiment is used as the baseline of context study.

A more in-depth study of context detection is performed by *Ex M2* and *Ex M3*, where they pass the detection of MVE contexts to another unsupervised learning system. The following two subsections discuss how to find the right representation of contexts through either geometric descriptors, or physics-related functions. Once the representation factors are chosen, K-means clustering algorithm [57] is applied to construct a number of MVE clusters, each of which would become a separate learning context. Clustering algorithms are widely used in unsupervised learning, where they are presented with a set of data instances that must be grouped according to some notion of similarity. In the case

of K-means, the Euclidean distance of (potentially, high dimensional) features is used as the measure of similarity to group the data into K clusters. The algorithm is unsupervised, as it has access only to the set of features describing each object, but it is not given any information (e.g., cluster labels) as to where each of the object should be placed. The result of K-means clustering is K cluster centers, identified by coordinates in the feature space. To determine K, we did a grid search between 90 and 110 and evaluated the purity in terms of VF variance within a cluster. During testing, the closest cluster to which a test sample should be assigned is identified by computing its distance with all the cluster centers and selecting the one with the minimum distance.

Context Detection with Connected Components

In *Ex M2*, we design seven macroscale descriptors to discriminate MVE-wise statistics: volume fraction, number of connected components, equivalent radius of the largest, smallest and average connected component, components' nearest surface distance, as well as nearest center distance. Besides the volume fraction, all other MVE-level feature descriptors depend on the identification of connected components in a MVE, the details of which are discussed next.

Connected-component labeling is used in computer vision to detect connected regions in binary digital images. A connected region or cluster is defined by pixels linked together by a defined rank, or level of neighbors in our setup. In terms of pixel connectivity in 3D, 6-connected, 18-connected, and 26-connected are commonly used to define a connected component. They correspond to a coverage of voxels within a squared Euclidean distance of 1, 2, and 3, to be regarded as connected. The concept of connectivity gets looser from 6-connected to 26-connected.

Obtaining connected components from our MVE takes additional effort, considering the assumed periodic nature of the microstructure. The voxels at a border surface are considered to be connected with the voxels on the opposite surface, and the usual ways of connected component labeling with pixel traversal would give a biased result. We compute the unbiased estimate of the number of connected components by following the following procedure:

1. Given an MVE, obtain the usual connected component number c_0.
2. Make a copy of the MVE and have the pair concatenated in three ways: along a border side, along a border line, and along a border node. Denote each concatenation type with an index i, $i = 1, 2, 3$. Within each type, there may exist multiple variations, indexed by j, $j = 0, 1, 2, \ldots$. Figure 3 shows for different concatenation variations of an MVE, using a real data instance.
   (a) Compute the number of connected components for each of the double-MVE structure $c_{i,j}$.
   (b) Calculate $d_{i,j} = 2 * c_0 - c_{i,j}$, which represents the number of mergeable or overcounted clusters for the given concatenation variation.
   (c) Sum up $d_{i,j}$ for each type of concatenation, i.e., $d'_i = \sum_j d_{i,j}$.
3. The final estimate of the unbiased count of connected components is $c_{\text{unbiased}} = c_0 - d'_1 + d'_2 - d'_3$. The rationale behind it is to try to have a correct count of connected components when all periodicity scenarios are considered. Following the decreasing order of side - line - node, most of the potential components to be merged with others at a lower order should also have been identified at a higher order. For example, a cluster appearing at a border node should have most likely been seen at the associated border line(s) as well. Therefore, to obtain a reasonable estimate of the unbiased count, we subtract the number of components that are miscounted as singulars in c_0 but turn out to be mergeable at the side concatenation level (d'_1), and then add back those at the line concatenation level (d'_2) that might been incorrectly subtracted at the previous level, and finally subtract those at the node concatenation level (d'_3) that have possibly been incorrectly added at previous level. It is important to note that this is simply an estimate of the actual count and may not be exactly accurate in special cases of odd-shaped clusters at MVE boundaries that would only be detected at lower concatenation levels, but nonetheless significantly corrects for the bias introduced by simply counting the clusters in a single MVE and ignoring the boundary conditions. Since we intend to only use it as a chosen input feature for the machine learning algorithm, it is fine to use an estimate for our purposes.

Note that if 6-connectivity is used the latter two concatenation forms $c_{2,*}$ and $c_{3,*}$ become irrelevant. Similarly 18-connectivity will render $c_{3,*}$ irrelevant. In our study, 18-connectivity is used, where neighbor positions up to level 2 are considered connected to the center.

After connected component labeling and obtaining the number of connected objects, we traverse through each object to find its equivalent radius, its nearest surface distance, and center distance towards other objects. The maximum, minimum, and average of the equivalent radius, and the minimum of the nearest surface distances, as well as the minimum of the nearest center distances, are used together as the set of macro features of the MVE.

Clustering algorithms take these features to determine the grouping, based on training data. As described before, the assignment of testing MVEs to clusters is determined by the distance to the center of cluster.

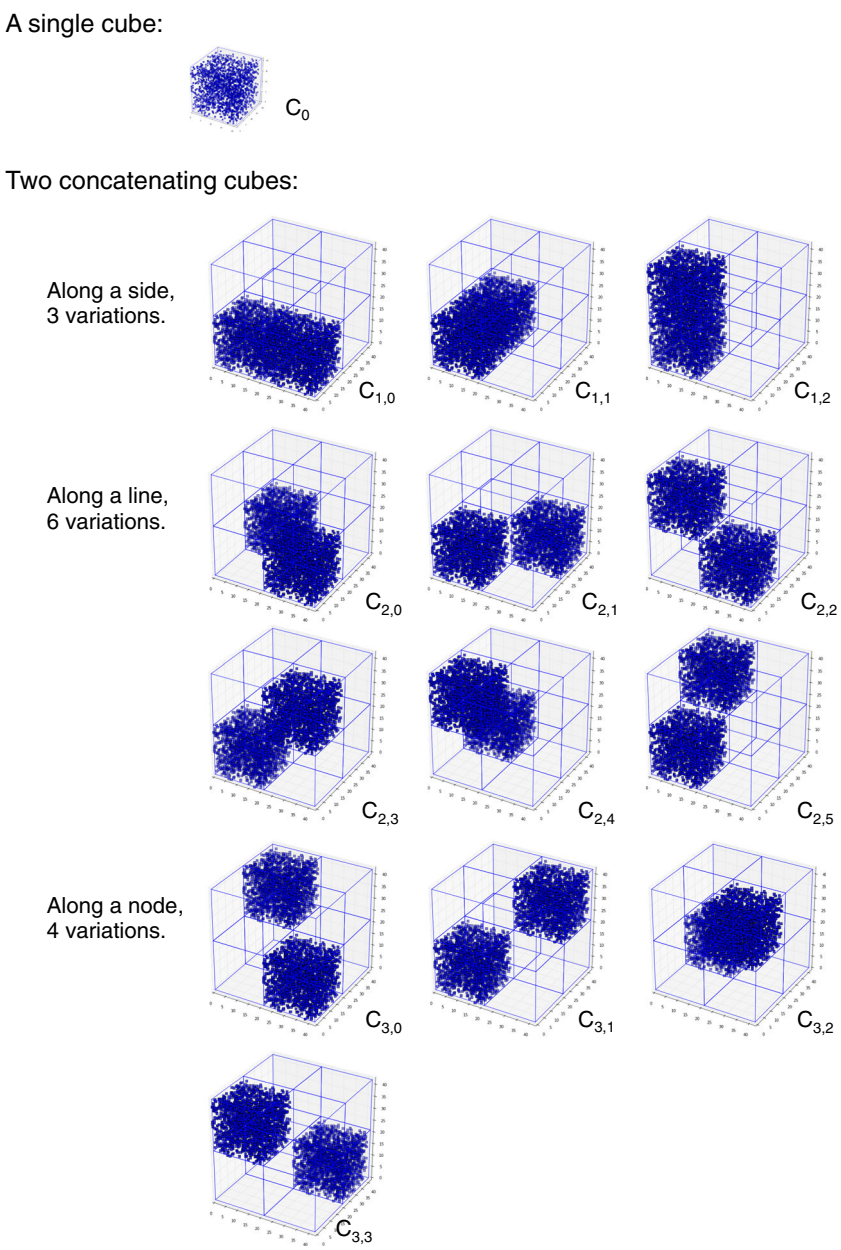

Fig. 3 Cube concatenation for computing connected components (Color figure online)

Context Detection with Pair Correlation Functions

Ex M3 further extends the automation of context detection by avoiding the need for manually defining descriptors. The design follows the concept of having (1) a more complete representation of microstructures and (2) a generalized statistical compression of such representation. This concept is carried out with the practice of pair correlation function (PCF) and principal component analysis (PCA).

Pair correlation functions $P_r^{hh'}$ denote the probability of finding discrete local states of h and h' separated by a distance r. PCF is widely used in the microstructure quantifications [58, 59], microstructure reconstructions [60], and estimation of effective properties [61]. For the case of $h = h'$, the value corresponding to distance $r = 0$ gives the volume fraction of the selected phase (since only one distinct local state can exist in a spatial voxel). The number of unique distances that exist in the $21 \times 21 \times 21$ periodic MVEs used in this study is 179. Even though there are finite number of dimensions in this representation, the PCFs decay quickly to a value specific to the selected MVE and fluctuate around it.

Even though PCF represents the microstructure in an efficient and accurate way, the number of dimensions is too large for building computationally efficient structure-property linkages. PCA is employed to decrease

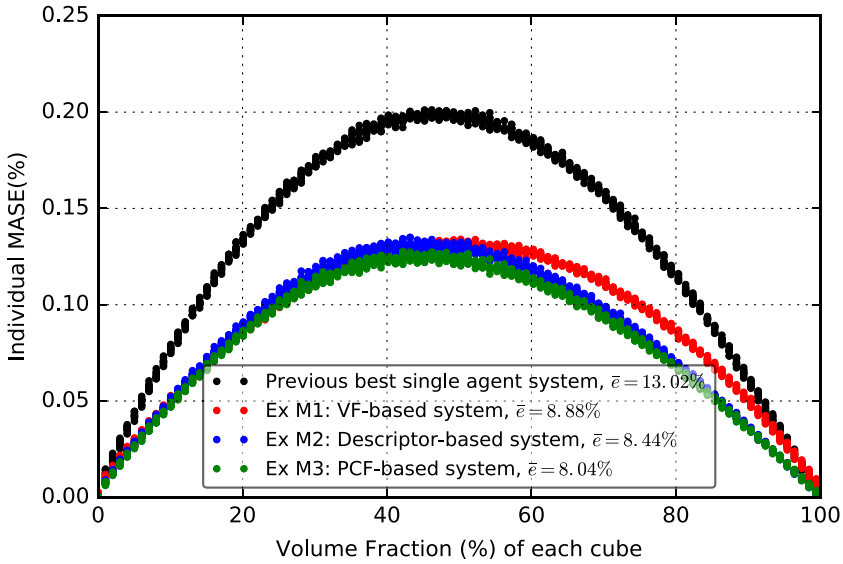

Fig. 4 Individual MASE attained by the three proposed context detection systems for each of the 1000 MVEs in test set, along with the MASE for the best single agent system [7] (Color figure online)

the dimensions of the microstructure representation. In our prior work, PCA was successfully implemented for both dimensionality reduction [23, 25], microstructure classification [9, 10], process-structure-property linkages [28, 30]. PCA performs the dimensionality reduction by transforming the coordinate system where the data is represented in such a way that the components of the new coordinate system are ordered from highest variance to lowest variance.

Result Analysis

The quality of learning systems is gauged by their performance on new validation data (unseen in calibration), in this case, the set of 1000 validation MVEs. The prediction error e (defined in "Dataset Details") for each individual MVE is shown in Fig. 4, with regards to the volume fraction. Essentially, in Fig. 4, at each volume fraction value on x-axis, there are 10 data points. The same pattern across all systems is observed: a parabolic dependence of the error value is observed on the volume fraction. In other words, the models developed have the highest errors for MVEs with volume fractions close to 0.5.

The average prediction error $\bar{e}$ across all 1000 MVEs for different systems is shown in Table 3. By comparing with the previous best single-agent learning system presented in [8], we observe that the proposed multi-agent learning systems improve the prediction performance in terms of test error, by as much as 38% ($1 - 8.04/13.02$). All three multi-agent systems proposed in this paper achieve a test MASE around 8%, which validates that the concept of context extraction is constructive in producing more accurate prediction systems. It should be noted that the error measures of "M3" can be further reduced by using more rigorous microstructure quantification methods [9, 25, 28, 30]. It is also rather trivial to impose a multi-agent layer onto any existing prediction system. Computationally, analyzing how data samples naturally form clusters is not very demanding and is only required one-time (before training takes place). The efficiency depends on the size of features that go into the clustering algorithm. As a result, in the current setting M1 is the fastest because only one feature (VF) is used. Computational times of M2 and M3 are similar.

The prediction output of a center slice of a randomly selected MVE (shown is the 500th MVE, VF=50%) is illustrated in Fig. 5. The difference between a ground truth slice (top one, annotated "FE") and a predicted slice (bottom ones, annotated "M1," "M2," "M3," respectively) is almost indistinguishable. In addition to a pixel-by-pixel comparison of slices, alongside the slices, we also have a statistical comparison between the ground truth and predicted strain distributions, over the whole MVE. Distributions are separated in color with regards to the phase (h shown in the plots).

Table 3 Data experiment results. Different context detection strategies are presented with training and testing error (the lower the better), along with the previous best single-agent system in [8]

Experiment	Context division based on	Training MASE (%)	Testing MASE (%)
Previous best	–	7.17	13.02
Ex M1	VF	6.21	8.88
Ex M2	Descriptors	5.70	8.44
Ex M3	correlation function	5.73	8.04

Testing error is regarded as a gauge of the prediction system quality

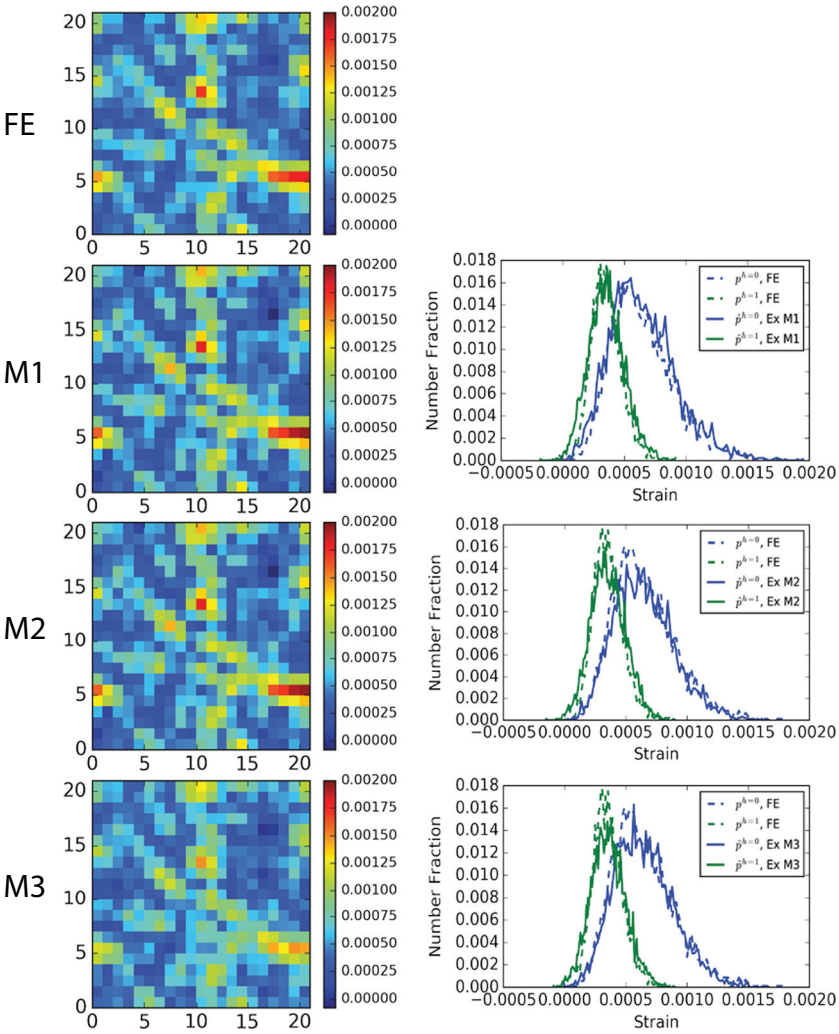

Fig. 5 Comparison of strain field predictions. A comparison of FE and statistical model predictions of the strain fields, $\hat{p}$, in an example MVE with the volume fraction of 50%. The models are those developed in Ex M1, M2, and M3. Strain fields on a center slice are compared (Color figure online)

Conclusions

In spite of the tremendous popularity and interest in the use of data science and informatics approaches to capturing PSP linkages in advanced hierarchical materials, the design of predictive models still remains far from optimal. Problems such as overfitting, performance discrepancy between training, and testing sets persist. A major reason for these problems is that it is often difficult to identify an accurate representation of structure-property relationship. Even though there are numerous efforts towards learning such a representation, they are often conducted in a flat manner, i.e., simply treating every pair of structure-property in the training data as if they come from the same distribution. Such treatment is often insufficient for complex systems, as samples in training data could actually emerge from many causal environments, which we define as *contexts*. The challenge here is that the information about which context each data sample is most likely to be associated with is hidden and hard to infer. However, such multi-contextual challenge of modeling unknown distributions in data has to be addressed in order to achieve accurate modeling. Otherwise, the common practice of building one flat classifier can only address relatively simple systems.

In this paper, we developed and demonstrated a novel multi-agent learning framework that breaks the large-scale prediction problem into self-contained subproblems, intending to connect macroscale characteristics of microstructure MVEs to microscale characteristics of each localized voxel. We evaluated three strategies of learning microstructure similarity: (1) with only volume fraction, (2) with macroscale features designed from observing connectivity, and (3) pair correlation functions. With each of these methods, a two-level learning scheme was implemented. Firstly, the MVE-level learning extracts high-level morphology information of microstructures and attempt to place together those that are homogeneous in morphology in one subproblem. Then, another set of microscale features

look at local details at each specific spatial location. The two-scale modeling of microstructure information has boosted the predictive modeling performance vigorously. It is clear from these trials that context-aware machine learning strategies are an important toolset for establishing high value, low computational cost, PSP linkages in complex hierarchical material systems.

Acknowledgements All authors gratefully acknowledge primary support for this work from AFOSR award FA9550-12-1-0458. Additionally, partial support is also acknowledged from the following grants: NIST award 70NANB14H012; NSF award CCF-1409601; DOE awards DESC0007456, DE-SC0014330; and Northwestern Data Science Initiative.

Author Contributions RL performed the data experiments and drafted the manuscript with help from YCY, ZY, SRK, and AA. ANC and AA provided the data mining expertise, and SRK provided domain expertise. AA supervised the overall design, development, and implementation of the proposed learning methodology. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Competing Interests The authors declare that they have no competing interests.

References

1. Agrawal A, Choudhary A (2016) Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater 4(053208):1–10
2. Kalidindi S, Medford AJ, McDowell DL (2016) Vision for data and informatics in the future materials innovation ecosystem. JOM 68(8):2126–2137
3. Panchal JH, Kalidindi S, McDowell DL (2013) Key computational modeling issues in integrated computational materials engineering. Comput-Aided Des 45(1):4–25
4. Ward L, Agrawal A, Choudhary A, Wolverton C (2016) A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater 2(16028)
5. Deshpande P, Gautham B, Cecen A, Kalidindi S, Agrawal A, Choudhary A (2013) Application of statistical and machine learning techniques for correlating properties to composition and manufacturing processes of steels. In: 2nd World Congress on Integrated Computational Materials Engineering (ICME), pp 155–160
6. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak J, Thompson A, Zhang K, Choudhary A, Wolverton C (2014) Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B 89(9):094104
7. Liu R, Kumar A, Chen Z, Agrawal A, Sundararaghavan V, Choudhary A (2015) A predictive machine learning approach for microstructure optimization and materials design. Scient Rep 5
8. Liu R, Yabansu YC, Agrawal A, Kalidindi S, Choudhary A (2015) Machine learning approaches for elastic localization linkages in high-contrast composite materials. Integr Mater Manuf Innov 4(1):1–17
9. Niezgoda SR, Yabansu YC, Kalidindi S (2011) Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Mater 59(16):6387–6400
10. Niezgoda SR, Kanjarla AK, Kalidindi S (2013) Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. Integr Mater Manuf Innov 2(1): 1–27
11. Gopalakrishnan K, Agrawal A, Ceylan H, Kim S, Choudhary A (2013) Knowledge discovery and data mining in pavement inverse analysis. Transport 28(1):1–10
12. Agrawal A, Deshpande PD, Cecen A, Basavarsu GP, Choudhary A, Kalidindi S (2014) Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integr Mater Manuf Innov 3(1):1–19
13. Yabansu YC, Patel DK, Kalidindi S (2014) Calibrated localization relationships for elastic response of polycrystalline aggregates. Acta Mater 81:151–160
14. Yabansu YC, Kalidindi S (2015) Representation and calibration of elastic localization kernels for a broad class of cubic polycrystals. Acta Mater 94:26–35
15. Agrawal A, Meredig B, Wolverton C, Choudhary A (2016) A formation energy predictor for crystalline materials using ensemble data mining. In: Proceedings of IEEE International Conference on Data Mining (ICDM), pp 1276–1279
16. Agrawal A, Choudhary A (2016) A fatigue strength predictor for steels using ensemble data mining. In: Proceedings of 25th ACM International Conference on Information and Knowledge Management (CIKM), pp 2497–2500
17. Gagorik AG, Savoie B, Jackson N, Agrawal A, Choudhary A, Ratner MA, Schatz GC, Kohlstedt KL (2017) Improved scaling of molecular network calculations: the emergence of molecular domains. J Phys Chem Lett 8(2):415–421
18. Fullwood DT, Niezgoda SR, Adams B, Kalidindi S (2010) Microstructure sensitive design for performance optimization. Progress Mater Sci 55(6):477–562
19. Liu R, Agrawal A, Chen Z, Liao W-K, Choudhary A (2015) Pruned search: a machine learning based meta-heuristic approach for constrained continuous optimization. In: Proceedings of 8th IEEE International Conference on Contemporary Computing (IC3), pp 13–18
20. Suh C, Rajan K (2009) Invited review: data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure–property relationships. Mater Sci Technol 25(4):466–471
21. Rajan K (2005) Materials informatics. Mater Today 8(10):38–45
22. Ward L, Liu R, Krishna A, Hegde V, Agrawal A, Choudhary A, Wolverton C (2016) Accurate models of formation enthalpy created using machine learning and voronoi tessellations. In: APS Meeting Abstracts
23. Choudhury A, Yabansu YC, Kalidindi S, Dennstedt A (2016) Quantification and classification of microstructures in ternary eutectic alloys using 2-point spatial correlations and principal component analyses. Acta Mater 110:131–141
24. Furmanchuk A, Agrawal A, Choudhary A (2016) Predictive analytics for crystalline materials: Bulk modulus. RSC Adv 6(97):95246–95251
25. Steinmetz P, Yabansu YC, Hötzer J, Jainta M, Nestler B, Kalidindi S (2016) Analytics for microstructure datasets produced by phase-field simulations. Acta Mater 103:192–203
26. Liu R, Ward L, Wolverton C, Agrawal A, Liao W-K, Choudhary A (2016) Deep learning for chemical compound stability prediction. In: Proceedings of ACM SIGKDD Workshop on Large-scale Deep Learning for Data Mining (DL-KDD), pp 1–7
27. Liu R, Agrawal A, Liao W-K, De Graef M, Choudhary A (2016) Materials discovery: understanding polycrystals from large-scale electron patterns. In: Proceedings of IEEE Big Data Workshop on

Advances in Software and Hardware for Big Data to Knowledge Discovery (ASH), pp 2261–2269
28. Yabansu YC, Steinmetz P, Hötzer J, Kalidindi S, Nestler B (2017) Extraction of reduced-order process-structure linkages from phase-field simulations. Acta Mater 124(1):182–194
29. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
30. Gupta A, Cecen A, Goyal S, Singh AK, Kalidindi S (2015) Structure–property linkages using a data science approach: application to a non-metallic inclusion/steel composite system. Acta Mater 91:239–254
31. Nguyen S, Tran-Le A, Vu M, To Q, Douzane O, Langlet T (2016) Modeling thermal conductivity of hemp insulation material: a multi-scale homogenization approach. Build Environ 107:127–134
32. Zhou X-Y, Gosling P, Pearce C, Ullah Z (2016) Perturbation-based stochastic multi-scale computational homogenization method for the determination of the effective properties of composite materials with random properties. Comput Methods Appl Mech Eng 300:84–105
33. Cruzado A, Gan B, Jiménez M, Barba D, Ostolaza K, Linaza A, Molina-Aldareguia J, Llorca J, Segurado J (2015) Multiscale modeling of the mechanical behavior of in718 superalloy based on micropillar compression and computational homogenization. Acta Mater 98:242–253
34. Fast T, Kalidindi S (2011) Formulation and calibration of higher-order elastic localization relationships using the MKS approach. Acta Mater 59(11):4595–4605
35. Landi G, Niezgoda SR, Kalidindi S (2010) Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel dft-based knowledge systems. Acta Mater 58(7):2716–2725
36. Landi G, Kalidindi S (2010) Thermo-elastic localization relationships for multi-phase composites. Comput, Mater, Contin 16(3):273–293
37. Guo N, Zhao J (2016) 3d multiscale modeling of strain localization in granular media. Computers and Geotechnics
38. Seko A, Maekawa T, Tsuda K, Tanaka I (2014) Machine learning with systematic density-functional theory calculations: application to melting temperatures of single-and binary-component solids. Phys Rev B 89(5):054303
39. Bhadeshia H, Dimitriu R, Forsik S, Pak J, Ryu J (2009) Performance of neural networks in materials science. Mater Sci Technol 25(4):504–510
40. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G (2003) Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 91(13):135503
41. Fischer CC, Tibbetts KJ, Morgan D, Ceder G (2006) Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 5(8):641–646
42. Da Silva BC, Basso EW, Bazzan AL, Engel PM (2006) Dealing with non-stationary environments using context detection. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, pp 217–224
43. Kalidindi S, Niezgoda SR, Landi G, Vachhani S, Fast T (2010) A novel framework for building materials knowledge systems. Comput, Mater, Contin 17(2):103–125
44. Fast T, Niezgoda SR, Kalidindi S (2011) A new framework for computationally efficient structure–structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models. Acta Mater 59(2):699–707
45. Kröner E (1986) Statistical modelling. Springer, Netherlands, pp 229–291
46. Kröner E (1977) Bounds for effective elastic moduli of disordered materials. J Mech Phys Solids 25(2):137–155
47. Fullwood DT, Niezgoda SR, Kalidindi S (2008) Microstructure reconstructions from 2-point statistics using phase-recovery algorithms. Acta Mater 56(5):942–948
48. Hibbitt Karlsson Sorensen (2001) ABAQUS/standard User's Manual, vol 1. Hibbitt, Karlsson & Sorensen, Providence, RI
49. Kalidindi S, Landi G, Fullwood DT (2008) Spectral representation of higher-order localization relationships for elastic behavior of polycrystalline cubic materials. Acta Mater 56(15):3843–3853
50. Al-Harbi HF, Landi G, Kalidindi S (2012) Multi-scale modeling of the elastic response of a structural component made from a composite material using the materials knowledge system. Modell Simul Mater Sci Eng 20(5):055001
51. Panait L, Luke S (2005) Cooperative multi-agent learning: the state of the art. Auton Agent Multi-Agent Syst 11(3):387–434
52. scikit-learn: Machine Learning in Python. http://scikit-learn.github.io/. [Online; accessed August 2015]
53. Garmestani H, Lin S, Adams B, Ahzi S (2001) Statistical continuum theory for large plastic deformation of polycrystalline materials. J Mech Phys Solids 49(3):589–607
54. Saheli G, Garmestani H, Adams B (2004) Microstructure design of a two phase composite using two-point correlation functions. J Comput-aided Mater Des 11(2-3):103–115
55. Fullwood DT, Adams B, Kalidindi S (2008) A strong contrast homogenization formulation for multi-phase anisotropic materials. J Mech Phys Solids 56(6):2287–2297
56. Adams B, Canova GR, Molinari A (1989) A statistical formulation of viscoplastic behavior in heterogeneous polycrystals. Textures Microstruct 11:57–71
57. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol 1, CA, USA, pp 281–297
58. Torquato S (2002) Statistical description of microstructures. Ann Rev Mater Res 32(1):77–111
59. Torquato S (2002) Random heterogeneous materials: microstructure and macroscopic properties, vol 16. Springer, New York
60. Liu Y, Greene MS, Chen W, Dikin DA, Liu WK (2013) Computational microstructure characterization and reconstruction for stochastic multiscale material design. Comput-Aided Des 45(1):65–76
61. Øren P-E, Bakke S (2002) Process based reconstruction of sandstones and prediction of transport properties. Transp Porous Media 46(2-3):311–343